Reproducible Statistical Analysis in Microarray Profiling Studies

U. Mansmann[1], M. Ruschhaupt[1, 2], W. Huber[3]
[1]IBE, University of Munich, Muenchen, Germany
[2]Division of Molecular Genome Analysis, German Cancer Research Center, INF 580, Heidelberg, Germany
[3]EBI, EMBL, Cambridge, UK

Summary

Objectives: Microarrays are a recent biotechnology that offers the hope of improved cancer classification. A number of publications presented clinically promising results by combining this new kind of biological data with specifically designed algorithmic approaches. But, reproducing published results in this domain is harder than it may seem.
Methods: This paper presents examples, discusses the problems hidden in the published analyses and demonstrates a strategy to improve the situation which is based on the vignette technology available from the R and Bioconductor projects.
Results: The tool of a compendium is discussed to achieve reproducible calculations and to offer an extensible computational framework. A compendium is a document that bundles primary data, processing methods (computational code), derived data, and statistical output with textual documentation and conclusions. It is interactive in the sense that it allows for the modification of the processing options, plugging in new data, or inserting further algorithms and visualizations.
Conclusions: Due to the complexity of the algorithms, the size of the data sets, and the limitations of the medium printed paper it is usually not possible to report all the minutiae of the data processing and statistical computations. The technique of a compendium allows a complete critical assessment of a complex analysis.

Keywords

Classification, reproducibility, microarray profiling studies, statistical computation

Methods Inf Med 2006; 45: 139–45

1. Introduction

Microarray technology allows for the simultaneous measurement of thousands of transcripts within a homogeneous sample of cells [1]. It is of interest to relate these expression profiles to clinical phenotypes to improve the diagnosis of diseases and prognosis for individual patients [2]. A number of publications presented clinically promising results by combining this new kind of biological data with specifically designed algorithmic approaches. A selection out of these papers will be discussed [3-6] with respect to different aspects of reproducibility.

The most evident aspect of reproducibility is that of reproducing a calculation on the same data. Reproducing published results in the domain of microarray profiling studies is harder than it may seem. As an example we look at a study of van ’tVeer et al. [3] which was reanalyzed by Tibshirani and Efron [7]. Both state in their paper: *We re-analyzed the breast cancer data from van ’t Veer et al. ... Even with some help of the authors, we were unable to exactly reproduce this analysis.*

Reproducing the predictive accuracy of a genomic profile on an independent data set is the basic prerequisite in algorithmic modeling [8]. Hardly any paper on microarray profiling studies meets this criterion. An exception is presented by Bullinger et al. [6] who use a learning and a testing sample for the profile. Mostly, the authors report surrogates of the predictive accuracy derived from the learning sample by cross-validation strategies (CV). Amboise and McLachlan [9] show that the popular leave-one-out cross-classification method results in an over-optimistic predictive accuracy and they demonstrate the need to implement appropriate cross-validation strategies to get reproducible predictive accuracy for a genomic profile.

We do not know any examples where classification results gained with one microarray technology and a special algorithm was reproduced using an alternative microarray platform and algorithm. Papers with diverging results on profiles for the prognosis of tumor recurrence for breast cancer patients are [3, 4]. How does this observation relate to the idea of a common underlying disease process? Should profiles have something common which is developed for the same disease? Is it of significance if they do not?

Profiles gain clinical acceptability if they reproduce established biological knowledge for the problem under study. Chang et al. [5] present a profile to discriminate between responder and non-responder to docetaxel, a chemotherapy developed for adjuvant or neo-adjuvant use in patients with breast cancer. Their classifier does not include genes that have previously been associated with taxane resistance [10].

Reproducibility is an important issue for reviewing a submitted microarray profiling study. How can you reduce the burden on the reviewer and allow for a timely and competent response? Making data available is not enough. The analysis is complex and depends on many parameters which are not known to the reviewer. The classical solution which bundles the paper, the primary data and the code in a zip-file does not relieve the reviewer: reanalysis requires manual interaction, the code may not be totally clear. This may result in a tendency to accept seemingly realistic computational results, as presented by figures and tables, without any proof of correctness [11].

How to judge the quality of a classification algorithm for a problem at hand? Some authors tabulate measures of performance for selected algorithms applied to selected data sets under selected preprocessing strategies [12, 13]. What can be learned from such numerical exercises for other problems? How can the researcher to gain fair and honest information on the methodological background she/he can use to set up a profiling study or to evaluate it correctly?

The confounding of algorithmic problems with biotechnology and biology creates a gordic knot. The paper applies the tool of a compendium as an interactive strategy to settle the algorithmic backbone of a profiling study and to derive reproducible results with a state-of-the-art methodology. Based on a reproducible calculation, one can proceed to the more interesting questions regarding the relationship between microarray platforms and signatures and biological interpretability. The paper presents an alternative to the static approach described in the last paragraph. An interactive document which allows reproducing, changing, or expanding a study solves the problems discussed before. Each step of the calculations with all related minutia can be verified. Alternative profiling strategies can be compared for a given data set. The dependence of a profiling strategy to specific data sets can readily be assessed.

A compendium is a document that bundles primary data, processing methods (computational code), derived data, and statistical output with textual documentation and conclusions. It is interactive in the sense that it allows the modification of processing options, plugging in new data, or inserting further algorithms and visualizations. We use the vignette technology available from the R and Bioconductor projects [14, 15] which is based on Sweave, a literate data analysis approach developed by Friedrich Leisch [16].

The paper discusses the reproducibility of a statistical calculation on a given data set. It relates to the problems with an efficient reviewing process of papers presenting results of microarray profiling studies, as well as how to assess critically published results to prepare future projects. The proposed tools are a starting point to explore how a different choice of the data processing technique affects the outcome of a microarray profiling study.

2. Examples and Questions

Two examples will be used to sharpen the questions related to reproducible statistical analysis in microarray profiling studies.

2.1 Example 1

Van 't Veer et al. [3] classify breast cancer patients after curative resection with respect to the risk of tumor recurrence. The study includes 78 patients. Forty-four patients had a good prognosis and did not suffer from a recurrence during the first five years after resection. Thirty-four patients had a bad prognosis and experienced a recurrence during the first five years after resection. Agilent microarray technology was used to quantify the transcripts probed by 24,881 oligonucleotides. Additional prognostic factors like tumor grade, ER status, PR status, tumor size, patient age, and angioinvasion were also documented. The authors were interested in developing a classifier based on the gene expression and to compare the relevance of the genomic signature to the prognostic value of standard clinical predictors.

They used the following algorithm to establish the signature which contains 70 genes:

1) Starting with 24,881 genes, filtering on fold-change and a p-value criterion reduced the number of relevant genes to 4936.
2) An absolute correlation of at least 0.3 between gene expression and group indicator (0, 1) resulted in a further reduction on 231 genes
3) Calculation of the 231 dimensional centroid vector for the 44 good prognosis cases.
4) Correlation of each case with this centroid is calculated; a cut-off of 0.38 is chosen to exactly misclassify three with poor prognosis
5) Case is classified as good prognosis if correlation calculated for some number n of genes ($1 \leq n \leq 231$) with the centroid vector is ≥ 0.38; otherwise the case is classified to the *bad prognosis* group.
6) Starting with the top 5, 10, 15, ... genes, the classification procedure is carried out with leave-one-out cross-validation in order to pick the optimal number of genes →70

Based on this algorithm, van 't Veer et al. achieved a correct classification for 26 of 44 patients without recurrence and for 31 of 34 with recurrence. Tibshirani and Efron [7] tried to reproduce this results and report: *Even with some help of the authors, we were unable to exactly reproduce this analysis.* The differences between both calculations were not dramatic, but the example shows that algorithmic reproducibility is not a trivial issue and may have subjective elements. The algorithm is quite popular and is also used in [5] and other papers.

The van 't Veer examples raises the following questions: Why was a heuristic classification algorithm chosen and not a standard algorithm from machine learning? Are its computational aspects well understood? How important is the choice of the parameters for correct classification? Is the result easy to interpret? What justifies the popularity of the algorithm? Is the leave-one-out CV strategy appropriate? What is the effect of other CV strategies on the classification result?

2.2 Example 2

Huang et al. [4] investigated primary tumor samples from 52 patients with breast tumors and 1-3 positive lymph nodes. Eighteen patients had a recurrence within three years after surgery, and 34 patients did not. The authors concluded that they could predict tumor recurrence with misclassification rates of 2/34 and 3/18, respectively.

The authors presented a tree classifier with split decisions based on an a posteriori distribution over all possible splits. A description is available in the form of a technical report on the webpage of one of the authors. Due to ambiguities we did not succeed in translating the statistical ideas into software with which we could reproduce the analysis. More importantly, there is no *official* software version of the algorithm.

The authors perform two-dimension reduction steps before they apply the Bayesian tree classifier. First, the 12,625 probe sets on the HGU95Av2 Affymetrix GeneChips are reduced to 7030 by excluding probe sets with maximum intensities below 2^9 and a low variability across the samples. The second reduction creates 496 *metagenes* out of the 7030 probe sets by performing k-means clustering and using the first principal component of each cluster as an expression measure for the *metagene.*

As in example 1, this study is concerned with the prognosis of tumor recurrence of breast cancer patients after curative resection of the tumor. The authors state that they could not find any of the 70 genes used in the classifier of van 't Veer et al. in any of the metagenes which come up in the Bayesian tree classifier.

The Huang example raises the following questions: Why is it impossible for us to reproduce the good classification result? Why is the classification based on the idiosyncratic method so good? The preprocessing is not part of the CV loop; what influence might this have on the misclassification rate? How is it possible to disentangle computation, technology, and biology? Can we find a link between the Huang and van 't Veer classifiers?

3. The Compendium

Publications on microarray profiling studies often present one new microarray data set and one new classification method. Is it necessary to develop an idiosyncratic classification approach for each specific data set? Which classification result could be achieved with standard approaches? What loss in accuracy has to be traded for a rise in interpretability? How can I use new data to validate former results? If validation creates discrepancies, how is it possible to assess the contribution made by algorithmic aspects: error in implementation or no success with validation? Do I have sufficient instructions and details to reproduce the method under validation in an exact way? How dependent is the classification result on the method used? What can be learned from a published profiling study for future projects? To answer these and other questions we introduce the compendium of computational diagnostic tools (CCDT)

3.1 Compendium for Computational Diagnostic Tools – CCDT

The compendium is an interactive document that calculates the misclassification rate (MCR) for different classification methods. The validation strategy is fixed but may be modified by setting specific parameters. Therefore, outer and an inner cross-validation is mandatory: the inner CV tunes the algorithm-specific parameters (also including parameters for the preprocessing) and the outer CV estimates the misclassification rate. We see the preprocessing as part of the classification algorithm. The CV strategy is sketched in Figure 1.

Things that can be changed easily are: Classification methods, preprocessing steps, parameters for classification method and validation, and the data set.

The compendium allows combining guidelines with software, and embedding good statistical analysis in a text, which is accessible for medical or biological researchers. It is a document that bundles primary data, processing methods (computational code), derived data, and statistical output with the textual documentation and concludes it. Especially, it contains specific tools to represent results. The inclusion of an implemented classification method is via a wrapper. Writing a wrapper function for new classification algorithms is simple. So far, five classification approaches are already implemented: Shrunken centroids (PAM) [18], support vector machines (SVM) [19], random forest [20], general partial least squares, and penalized logistic regression [21]. The compendium compHuang and related material can be found as a package for the statistical language R at http://www.bioconductor.org/packages/bioc/1.7/src/contrib/html/MCRestimate.html. The compendium does not implement new-algorithms but offers a framework to apply existing implementations of classificational-gorithms in a correct way.

3.2 Static versus Interactive Approaches

The answer to questions like: *Which classification result could be achieved with standard approaches?* or *What loss in accuracy has to be traded for an increase in interpretability?* is generally given in a paper

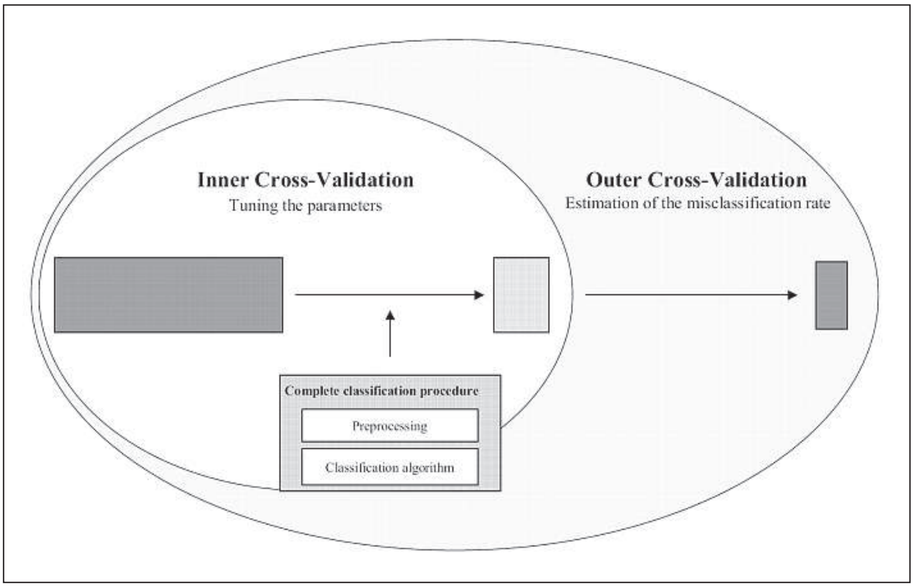


Fig. 1 The cross-validation strategy (CV – cross validation, MCR – misclassification rate)

which compares N different classification algorithms (C) and M pre-processing (P) strategies on K different data sets (D). The N × M × K performance measures are tabulated and discussed. Dudoit et al. [12] published such a study which is extended by Lee et al. [13]. It is difficult to use their results for guidance because they certainly do not implement all algorithms of interest, the pre-processing strategies may change, and the data will become irrelevant when a new generation of microarrays is introduced.

Therefore, it may be wise to replace the static by an interactive approach by offering the machinery which allows the researcher herself to perform such a study on the algorithms and pre-processing strategies of interest together with relevant data.

The compendium offers different levels of interactivity. It can be used to produce a textual output comparable with the static approach. On an intermediate level one interacts with the compendium by specifying parameters and data sets. For example, one could change the kernel of a support vector machine by simply changing the parameter poss.pars:

```
>poss.pars = c (list (cost = cost.range,
kernel = “linear”), poss.k)
```

This is the level of sensitivity analyses or of comparing the performance of implemented algorithms on different data sets. The advanced level of interaction consists in introducing new ideas like new classification algorithms or new tools for the presentation of the results. Writing wrapper functions for new classification methods is simple. The following example shows a wrapper for diagonal discriminant analysis. The essential part is the specific definition of the predict.function by user-specific needs and ideas:

```
>DLDA.wrap = function(x, y, pool = 1, ...) {
+ require(sma)
+ predict.function = function(testmatrix) {
+ res = stat.diag.da(ls = x, as.numeric(y),
testmatrix, pool = pool)$pred
+ return(levels(y)[res])
+ }
+ return(list(predict = predict.function, info
= list( )))
+ }
```

3.3 Revivable Documents

The standard to make a statistical analysis reproducible consists in bundling the report together with the data and code in an archive (zip) or on a website. With this practice, extensive manual interaction is needed to reproduce results. Sometimes the code may be too cryptic to be fully understandable.

Literate statistical programming was introduced by Carey [22]. This method for statistical practice suggests that documentation and specification occur at the same time as statistical coding. It combines code and documentation in a source file that can be woven into a description of the process, algorithms, and results obtained from the system and tangled into the actual code. The two primary steps in constructing the results from the literate document are *tangling*, which produced unformatted code files that can be compiled or evaluated using a language interpreter; and weaving, which produces formatted documentation files to be read by humans [11].

Recent tools [23] allow going one step further: including code for figures and tables instead of the actual graphs and numbers. The source contains no numerical or graphical results; only the codes needed to obtain these documents are revivable in the sense that they can be automatically updated whenever data or analysis change. Code for the analysis is embedded into a manuscript, which is then evaluated, and both code and/or the corresponding output go into the final document.

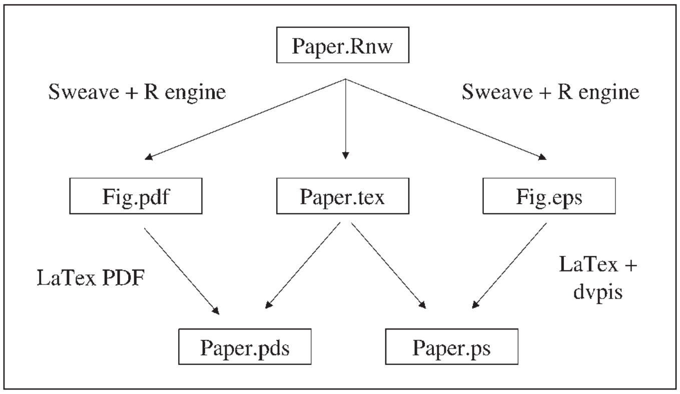


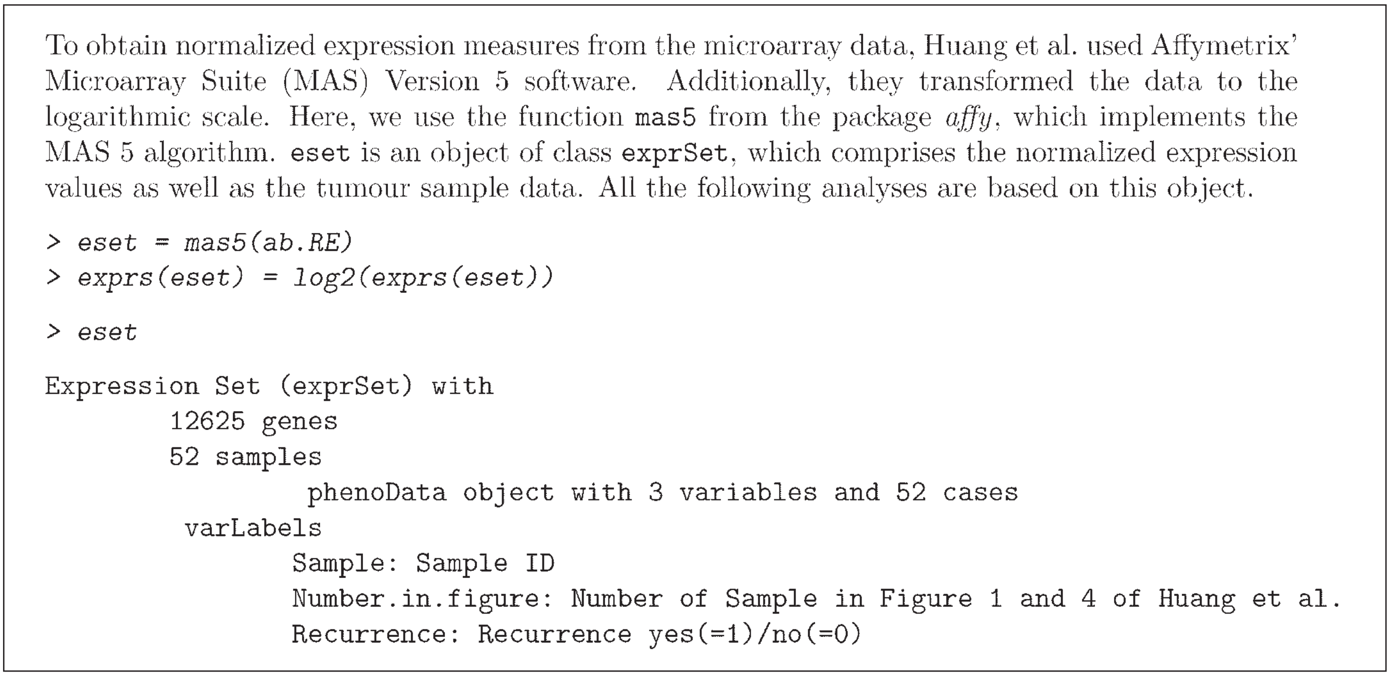
Fig. 2
The working principle of Sweave

3.4 Sweave

Sweave [16] is a specific approach for generating a dynamic report. We use Sweave as the technology for the compendium. It mixes S (R) and LaTex in a sequence of code and documentation chunks in an Rnw file. It uses S (R) for all tangling and weaving steps and hence has very fine control over the S (R) output. Options can be set either globally to modify the default behavior or separately for each code chunk to control how the output of the code chunks is inserted into the LaTex file.

The working principle of Sweave is sketched in Figure 2 and demonstrated in the following example. The parts between << ... >> = ... @ describe the code chunks which will be evaluated by S (R) and whose results will be woven, if required, into the output (again a LaTex of postscript document) or only available for later evaluation steps. The paragraphs between the code chunks will be processed as LaTex code for the output document.

To obtain normalized expression measures from the microarray data, Huang et al. used Affymetrix’ Microarray Suite (MAS) Version 5 software. Additionally, they transformed the data to the logarithmic scale. Here, we use the function \Rfunction{mas5} in the \Rpackage{affy} library. \Robject{eset} is an object of $class exprSet$, which comprises the normalized expression values as well as the tumor sample data. All the following analyses are based on this object.



To obtain normalized expression measures from the microarray data, Huang et al. used Affymetrix' Microarray Suite (MAS) Version 5 software. Additionally, they transformed the data to the logarithmic scale. Here, we use the function `mas5` from the package *affy*, which implements the MAS 5 algorithm. `eset` is an object of class `exprSet`, which comprises the normalized expression values as well as the tumour sample data. All the following analyses are based on this object.

```
> eset = mas5(ab.RE)
> exprs(eset) = log2(exprs(eset))

> eset

Expression Set (exprSet) with
        12625 genes
        52 samples
                 phenoData object with 3 variables and 52 cases
         varLabels
                Sample: Sample ID
                Number.in.figure: Number of Sample in Figure 1 and 4 of Huang et al.
                Recurrence: Recurrence yes(=1)/no(=0)
```

Fig. 3 Sweave output of example

```
%%normalizing the affy batches
<<normalizing, eval = FALSE >> =
Huang.RE <- mas5(affy.batch.RE)
exprs(Huang.RE <-log2(exprs(Huang.RE))
@
```

So we have the following expression set for our further analysis

```
<<show1>> =
Huang.RE
@
```

The Sweave output of this part is presented in Figure 3.

4. Results

The application of the compendium to data of microarray profiling study provides tools to answer crucial questions regarding the assessment of a new classification algorithm. A few aspects will be discussed.

A new classification algorithm can be implemented by writing the specific wrapper function. This is an easy exercise. Misclassification results can be calculated and compared to the results of a set of competing algorithms. The compendium presents classification results on the basis of the confusion matrix and individual classification. Individual classification based on a specific pre-processing strategy and classification algorithm can be graphically presented for the whole sample. Figure 4 shows a vote plot which, for each subject, presents the percentage of correct classification in the determined number of CV loops. The first 34 patients belong to the group with no tumor recurrence. The remaining 14 patients suffered a relapse.

A table giving a synopsis of all individual misclassifications can also be produced. Table 1 shows all subjects who are misclassified by at least one of the classification strategies under consideration in the data of Huang et al.

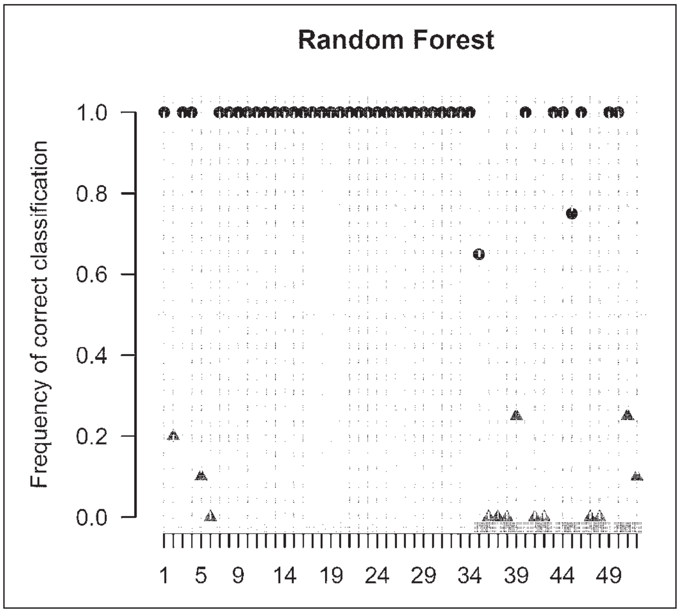


Fig. 4 Vote plot for classification result using random forest: reanalysis of the Huang et al. data [4]

Table 1 Reanalysis of Huang et al. data (RF – random forest, PAM – shrunken centoids, logReg – penalized logistic regression, SVM – support vector machine, M – using metagenes)

	RF-M	RF	PAM-M	PAM	PLR-M	PLR	SVM-M	SVM
6	×	×	×	×	×	×	×	×
36	×	×	×	×	×	×	×	×
37	×	×	×	×	×	×	×	×
38	×	×	×	×	×	×	×	×
41	×	×	×	×	×	×	×	×
42	×	×	×	×	×	×	×	×
47	×	×	×	×	×	×	×	×
48	×	×	×	×	×	×	×	×
39	×	×	×	×		×	×	×
51	×	×	×	×		×	×	×
45	×		×	×		×	×	×
35	×		×	×			×	×
52	×			×	×		×	×
2		×	×			×	×	
5		×			×	×		×
7		×						
40							×	
44							×	

Table 2 Misclassified 'recurrence' samples

	RF	PAM	logReg	SVM	RF-M	PAM-M	logReg-M	SVM-M
36	×	×	×	×	×	×	×	×
37	×	×	×	×	×	×	×	×
38	×	×	×	×	×	×	×	×
41	×	×	×	×	×	×	×	×
42	×	×	×	×	×	×	×	×
47	×	×	×	×	×	×	×	×
48	×	×	×	×	×	×	×	×
51	×	×	×	×	×	×	×	×
39	×	×	×		×	×	×	×
52	×	×			×	×	×	×
35		×			×	×	×	×
45		×			×	×	×	×

[4]. The individuals are ordered with respect to the number of classification strategies which lead to misclassification. Other presentations of the comparison are possible, for example, a scatter plot to contrast individual MCRs between the new and the standard approaches. This idea can also be implemented on the advanced level of interaction by writing the code for the respective figure.

The Huang strategy misclassifies two of the 34 patients with good prognosis and three of the 18 patients with bad prognosis. The standard algorithms give results on correct classification below 80%. Why is it impossible for us to reproduce the good classification result with standard algorithms? No implementation for the Bayesian classification tree (BCT) is available and thus no direct comparison is possible. The algorithm of the BCT is not available and its description in a technical report does not give explicit advice for its implementation into software. Therefore, we tried a sensitivity analysis by using the intermediate interaction with the compendium. First, we excluded the preproceeding form of the inner CV loop because the original paper [4] does not take care on the adjustment of the MCR for the pre-processing strategy. This did not improve classification quality in the expected way. Second, we reduced the data set to the 200 most discriminating genes. This introduces a huge selection bias. This was the only way we could come up with six to seven misclassifications, which is still worse as the results in the original paper. Can we trust the original result? Our analysis reminds us to be quite critical to the original claims.

Huang et al. [4] state that the genes found to be crucial for the classification are different from the genes found as crucial by van 't Veer et al. [3]. What is the reason for the missing link between the Huang and van 't Veer classifiers? How is it possible to disentangle computation, technology, and biology? The compendium takes care on the computational part. Additionally, the BCT and the van 't Veer classifier need to be implemented and wrapper functions for both classification algorithms have to be written. Both the wrapper and the respective preprocessing strategies will be applied to both data sets. Correlation between the genes used in the classifiers can be calculated and visualized by a checkerboard figure for each dataset. The checkerboard for genes of both classifiers on the same data set allows for the identification of genes with similar expression behavior and to study common biological aspects behind both classifiers. Comparing the checkerboards between both data sets gives hints on sample differences between both studies and discrepancies introduced by the different microarray technologies used.

5. Discussion

The literature on the induction of prognostic profiles from microarray studies is a methodological wasteland. Ambroise and McLachlan [9] describe the unthorough use of cross-validation in a number of high-profile published microarray studies. Tibshirani and Efron [7] report the difficulty in reproducing a published analysis. Huang et al. [4] present results with the potential to revolutionize clinical practice in breast cancer treatment but use an ideosyncratic statistical method which is fairly complex and neither easy to implement nor to obtain as software. A series of papers published in Nature [3], NEJM [6, 17], and The Lancet [4, 5] base their impressive results on classification methods which were developed ad-hoc for the problem at hand. The global picture looks like: the MCRs reported are of questionable clinical relevance, reanalysis of a paper does not support the strong claims they made, results are not reproducible with respect to computation, validation, or biology.

This situation has several implications: 1) It is nearly impossible to assess the value of the presented studies in terms of statistical quality and clinical impact. 2) Scientists looking for guidance to design similar studies are left puzzled by the plethora of methods. 3) It is left unclear how much potential there is for follow-up studies to incrementally improve on the results.

Our compendium offers a first approach to overcome these problems. It helps to give full account to the approach which is taken to understand a complex reality. The compendium is organized as a revivable document [23] and uses open source software: R, Cran, Bioconductor, LaTex [14-16]. The *classical solution* by bundling report, data, and code is not satisfying in setting microarray profiling studies. This way, reanalysis requires manual interaction and parts of code may not be understandable. A *revivable document* helps get computation straight, allows sensitivity analysis for a study, reanalysis of published studies, and may help to detect hidden confounding factors by using independent data from a different protocol. Getting the computation straight is necessary before successfully clarifying biological/medical questions.

One disadvantage of the compendium is the possibility of a *method bias*: overfitting by looking for the most successful classification strategy for the data at hand.

Certainly, it improves the statistician's task as the reviewer of microarray profiling studies. It reduces the **burden of the reviewer** and assists with timely response in the peer review process of a journal but also in the support for colleagues. The compendium creates a new communication platform between statisticians and biologists/ medical researchers. It helps to bring a study to publication standard.

Acknowledgment
This work was supported by the BMBF grant NGFN II FKZ 01 GR 0459.

References

1. Microarray special. Statistical Science 2003; 18: 1-117.
2. Simon R, Rademacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray classification. J Nat Cancer Inst 2003; 95: 14-8.
3. van 't Veer L, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Petersen HL, van de Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002; 415: 530-6.
4. Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, Bild A, Iversen ES, Liao M, Chen CM, West M, Nevins JR, Huang AT. Gene expression predictors of breast cancer outcomes. The Lancet 2003; 361: 1590-6.
5. Chang J, Wooten EC, Tsimelzon A, Hilsenbeck SG, Gutierrez MC, Elledge R, Moshin S, Osborne CK, Allred DC, O'Connell P. Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. The Lancet 2003; 362: 362-9.
6. Bullinger L, Döhner K, Bair E, Fröhling S, Schlenk RF, Tibshirani R, Döhner H, Pollack JR. Use of Gene-Expression Profiling to Identify Prognostic Subclasses in Adult Acute Myeloid Leukemia. NEJM 2004; 350: 1605-16.
7. Tibshirani RJ, Efron B. Pre-validation and inference in microarrays. Statistical Applications in Genetics and Molecular Biology 2002; 1 (1): Article 1.
8. Breiman L. Statistical Modelling: The Two Cultures. Statistical Science 2001; 16: 199-231.
9. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci 2002; 99: 6562-6.
10. Brenton JD, Caldas C. Predictive cancer genomics – what do we need? The Lancet 2003; 362: 340-1.
11. Leisch F, Rossini AJ. Reproducible statistical research. Chance 2003; 16: 41-6.
12. Dudoit S, Fridlyand J, Speed T. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. JASA 2002; 97: 77-87.
13. Lee JW, Lee JB, Park M, Song SH. An extensive comparison of recent classification tools applied to microarray data. Computational Statistics and Data Analysis 2004; (in press).
14. Ihaka R, Gentleman R. R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics 1996; 5: 299-314.
15. Gentleman R, Carey V. Bioconductor. R News 1996; 2: 11-6.
16. Leisch F. Sweave: Dynamic generation of statistical reports using literate data analysis. In: Härdle W, Rönz B (eds). Compstat 2002 – Proceedings in Computational Statistics. Heidelberg: Physika Verlag; 2002, pp 575-80.
17. van de Vijver MJ, He YD, van 't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002; 347: 1999-2009.
18. Tibshirani R, Hastie T, Narasimhan B, Chu G. Class prediction by nearest shrunken centroids, with application to DNA microarrays. Statistical Science 2003; 18: 104-17.
19. Vapnik V. The Nature of Statistical Learning Theory. New York: Springer; 1999.
20. Breiman L. Random Forests. Machine Learning Journal 2001; 45: 5-32.
21. Eilers PH, Boer JM, Van Ommen GJ, Van Houwelingen HC. Classification of Microarray Data with Penalized Logistic Regression. Proceedings of SPIE volume 4266: progress in biomedical optics and imaging, 2001; 2: 187-98.
22. Carey VJ. Literate Statistical Programming: Concepts and Tools. Chance 2001; 14: 46-50.
23. Sawitzki G. Keeping Statistics Alive in Documents. Computational Statistics 2002; 17: 65-88.

Correspondence to:
Ulrich Mansmann, PhD
Chair of Biometry and Bioinformatics
IBE, Medical School
LMU München
Marchioninistr. 15
81377 München
Germany
E-mail: mansmann@ibe.med.uni-muenchen.de
http://ibe.web.med.uni-muenchen.de